DOCUMENT RESUME

ED 208 060 TM 810 825

AUTHOR Baron, Joan Boykoff; And Others
TITLE Will the "Real" Proficiency Standard Please Stand Up?
PUB DATE 29 Apr 81
NOTE 44p.; Paper presented at the Annual Meeting of the New England Educational Research Organization (Lenox, MA, April 29, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Basic Skills; Comparative Analysis; *Cutting Scores; Educational Assessment; Elementary Secondary Education; *Methods; *Minimum Competency Testing; *Multiple Choice Tests
IDENTIFIERS Angoff Methods; Borderline Group Method; *Connecticut; Contrasting Groups Method; Nedelsky Method; Ninth Grade Proficiency Test

ABSTRACT

Connecticut's experience with four different standard-setting methods regarding multiple choice proficiency tests is described. The methods include Angoff, Nedelsky, Borderline Group, and Contrasting Groups Methods. All Connecticut ninth graders were administered proficiency tests in reading, language arts, and mathematics. As soon as final test forms were developed, the Connecticut State Department of Education (CSDE) began the process of setting standards. Four groups of subject matter experts participated as judges in the process using the Angoff and Nedelsky methods. For the Borderline and Contrasting Groups Methods, data were collected regarding proficient, borderline, or non-proficient students. General results showed that different techniques result in different cut scores and problem areas; thus, only one method should be used so implications of each cut score may be analyzed. Limitations concerning the methods are described, and recommendations are stated. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208060

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

WILL THE "REAL" PROFICIENCY STANDARD PLEASE STAND UP?

Joan Boykoff Baron, Douglas A. Rindone, Peter Prowda

The Connecticut State Department of Education
Box 2219
Hartford, Connecticut 06115
(203) 566-7232

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D Rindone

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of The New England Educational Research Organization

Symposium: Setting Standards in Achievement Tests of Basic Skills from Local School District and State Perspectives

Lenox, Massachusetts April 29, 1981

TM 810 825

With the recent proliferation of proficiency tests and the need of educators to make a variety of decisions based upon them, standard-setting methods have become extremely important. Glass (1978) has sensitized educators to the fact that all standard-setting methods are subjective and the experiences of standard-setters have confirmed it. Human judgment plays a central role in every standard-setting method.

The purpose of this paper is to describe Connecticut's experience with four different standard-setting methods. That experience clearly demonstrates the fact that different methods yield different results. Prospective standard-setters should be aware of these differences so as to choose the method or methods and their corresponding implementation strategies with care.

BACKGROUND

In June 1978, Connecticut joined 36 other states in mandating a policy of basic skills testing for its students. At three nonconsecutive elementary school grade levels, each district will select and administer a standardized norm referenced test. The results will be used to identify students in need of remedial assistance and to improve the general instructional program. Beginning in the School Year 1979-80, the State Board of Education will develop and administer a test in reading, language arts and mathematics to all students in the ninth grade. In March, 1980, the Connecticut State Department of Education administered to all ninth graders in Connecticut a proficiency test in reading, language arts and mathematics. The test was designed to measure those skills Connecticut students should possess after spending 8 years in the public school system. The content validity of the four subtests had been established in three ways: by committees of content experts; through responses to a survey sent to

5,000 Connecticut citizens including teachers, students, administrators, legislators, local school board members, representatives of business and industry and members of the general public; and by those citizens reviewing the skills at a series of six public meetings held throughout the state.

To ensure that the test items were, to the extent possible, not biased against any group on the basis of race, culture, language, or gender, a test bias committee was convened to examine (1) the items prior to the pilot testing and (2) the empirical data for each subgroup following the pilot test using transformed item difficulties (Angoff, 1972).

DESCRIPTION OF THE NINTH GRADE TEST AND SCORING

The mathematics test was composed of 65 multiple-choice items measuring 37 objectives or skills distributed among three domains of Computation, Concepts and Problem-Solving skills. The students had on hour in which to complete the test. Students received a total score and scores in the three domains.

The Basic Writing Skills in the Language Arts test was designed to assess the writing experience itself as well as eleven related language arts writing skills in the domains of Mechanics of Written Expression, Composing and Organizing Skills, and Library Skills for Writing Tasks. Accordingly, the test consisted of two parts: (1) an exercise requiring each student to write a passage based on personal experience (25 minutes), and (2) a 36 item multiple-choice test (40 minutes). The student receives a holistic score (2 to 8) on the writing portion. On the multiple choice test, the student receives a total score as well as scores on the 3 domains.

The Reading test was a 98 item Degrees of Reading Power (DRP) test marketed by The College Board. The 75 minute test consists of 14 nonfiction prose passages on a variety of topics. Each passage contains approximately 300 words and asks 7 questions. The passages are arranged by order of difficulty, beginning with very easy material and progressing to very difficult materials. Test items are formed

by the deletion of selected words in each passage. Each deleted word is indicated by an underlined black space. Five response options are provided to the students for completing the blank. The student receives a score based on the total number of correct responses as well as a DRP score which is an indication of the difficulty level of passages which the student would be able to comprehend.

METHODS AND PROCEDURES

Setting the Statewide Level of Expected Performance (SLOEP)

As soon as final test forms had been developed for each section of the March 1980 Ninth-Grade Proficiency Test, the Connecticut State Department of Education (CSDE) began the process of setting standards for the test. Education Evaluation and Remedial Assistance (EERA) Regulations mandated that a Statewide Level of Expected Performance (SLOEP) be established by January 1, 1980. Students whose scores are below the SLOEP will be further diagnosed and, if necessary, provided remedial assistance by the local or regional school board. School districts will be responsible for determining the specific needs of students whose scores are below the SLOEP and for developing programs to help those students.

Personnel from the Connecticut State Department of Education's EERA staff met with the Connecticut State Board of Education's EERA Advisory Committee to determine the procedures to be used for setting standards on the Connecticut test. The State Department staff made a proposal, based upon consultation with the Psychometrics Subcommittee,

which recommended using some combination of the four most commonly used procedures for setting standards on multiple-choice tests: (a) Angoff method, (b) Nedelsky method, (c) Borderline Group Method, and (d) Contrasting Groups method. The Advisory Committee debated the merits and drawbacks of each method and finally recommended that all four be used. However, because the regulations required that the SLOEP be set prior to administering the test, and since the Borderline Group and Contrasting Groups methods could not be utilized until after test scores from the March 1980 test had been made available, the EERA Advisory Committee recommended the following two steps:

(1) Use the Angoff and Nedelsky methods prior to January 1, 1980 to establish the statewide levels of expected performance for the March 1980 test administration.

(2) Use the Borderline and Contrasting Groups procedures after March 1980 to validate the SLOEP (set in Step 1) for future years.

The EERA Advisory Committee also assisted in determining the Statewide Levels of Expected Performance for the Writing exercise and for the reading test. The methods described above were not applicable to those tests. The following is a description of the standard-setting process for each portion of the Ninth-Grade Proficiency Test.

Mathematics and Language Arts multiple-choice tests. The Angoff and Nedelsky approaches to standard-setting both require the participation of subject-matter experts who know the capabilities and general performance levels of the student population and who are familiar with the curriculum in the schools. Four such groups of subject-matter experts, the majority of whom were teachers of ninth-grade students, participated as judges in the standard-setting process for the Connecticut mathematics and language arts multiple-choice tests. For each test, one group used the Angoff procedure and the other used the Nedelsky procedure. A brief summary of those methods is described here.

In both methods, the judges are required to imagine the minimally proficient student. For the Connecticut test, the minimally proficient student was defined as the one who could "survive" in his/her ninth-grade classes with his/her present level of proficiency in the basic skills of reading, language arts, and mathematics. Because the judges were chosen to be representative of urban, rural, and suburban schools, and because their experiences varied widely, each judge was asked to imagine that all of the 50,000 Connecticut ninth-graders were part of one "large" school district which now may be viewed as the "typical school district in the state." Thus, an attempt was made to select standards which could be applied in all school districts in the state.

The original subcommittees that had specified the skills to be tested and subsequently reviewed the items served as the Angoff Method standard-setters and used the Angoff method. They did so prior to

reviewing the pilot data so they did not know the p-values of the items when they set the standard. The members of the Angoff groups (nine judges for mathematics and ten for language arts)were asked to estimate for each test item, the probability that a minimally competent student would answer that question correctly. Or, phrased somewhat differently, they were asked, "In a group of 100 minimally competent (or proficient) students, how many of them would be expected to answer each question correctly?" The Mathematics and Language Arts committees using the Angoff method were trained together. All present made an estimate for a math item and then listened to all math experts discuss their rationales for high and low estimates. Then the procedure was repeated for a language arts item. The joint session took approximately one hour. Following this the two committees went into separate rooms to set the standards. Each judge made initial judgments for about ten items and then they discussed their probability estimates for each item. Following the discussion of each item, the judges made a revised probability estimate. It was the second set of responses that was then summed. An average of the sums was calculated to provide an estimate of the expected average score of a group of minimally proficient students with minimally acceptable performance

A second group of teachers was convened and trained to use the Nedelsky method. Those highly experienced teachers (an average of eighteen years of experience for the nine Mathematics teachers and an average of ninteen years of experience for the eleven Language Arts teachers) had not seen the test items prior to the standard-setting activity. The members of the Nedelsky group were asked a different question. "For each multiple-choice test item, identify the wrong answer choices that a minimally competent student would be able to recognize as wrong." Members of the group made initial and revised judgements. The latter were converted to probabilities

and summed. As with the Angoff method, an average of the sums was then calculated to provide an estimate of the expected average score of a group of students with minimally acceptable performance. Figure 1 contains the distribution of the individual judges' summed probabilities using the Angoff and Nedelsky methods on the mathematics and language arts tests.

Based upon a review of the Angoff and Nedelsky data, the CSDE recommended that the SLOEP for mathematics be set at 62 percent (41 items correct). In other words, students who answered 62 percent or more of the mathematics questions correctly would meet the expected performance standard. A SLOEP of 58 percent (21 items correct) was recommended for the language arts multiple-choice test.

Writing Sample. The writing sample was scored using a method known as "holistic scoring." When scored by the holistic method, student essays receive scores on a scale of 2 to 8, with 2 being the lowest score and 8 the highest.

The process of setting a SLOEP for the Connecticut writing sample involved the assistance of two groups, the EERA Advisory Committee and the Language Arts Subcommittee. Committee members, acting as judges, were asked to read a set of 18 papers selected from those which had been written for the pilot test. The judges were then asked to determine whether the writer (a) definitely needed remedial assistance, (b) definitely did not need remedial assistance, or (c) was on the borderline between needing remedial assistance and not needing it.

Mathematics

ANGOFF	SCORE	NEDELSKY
	48	x
	46	
	44	x
	42	xx
x	40	x
x	38	
xx	36	
x	34	x
	32	x

Mean of Ratings	37.6 41.5
Std Dev of Ratings	2.4 4.7
No of Items	65
Test Mean	49.3
Test Std Dev	11.2

Language Arts

ANGOFF	SCORE	NEDELSKY
	24	x
xx	22	
xxx	20	x
xx	18	xx
x	16	x
	14	x
	12	xx
	10	x

Mean of Ratings	21.0 17.5
Std Dev of Ratings	1.5 3.9
No of Items	36
Test Mean	29.6
Test Std Dev	5.7

Figure 1. The distributions of judges' scores using the Angoff and Nedelsky methods for the Mathematics and Language Arts tests.

After a brief training exercise, each judge rated the papers. Judges' ratings were then compared with the actual scores those papers had been given when scored holistically. Based upon their ratings, the two groups of judges agreed that papers which had received a summed score of 2 or 3 indicated a need for remedial assistance. The State Department, therefore, recommended as the SLOEP for the writing sample a holistic score of 4.

Reading. The process of setting a SLOEP for the Degrees of Reading Power Proficiency Test (DRP) involved the assistance of two groups, the EERA Reading Subcommittee and the EERA Advisory Committee.

The reading subcommittee used two procedures to establish a statewide level of expected performance. One sub-group examined the passages in the DRP asking themselves what was the most difficult passage which a ninth-grade minimally proficient student could be expected to read with 75% comprehension. The other sub-group examined lists of textbooks, commonly used in English and social studies classes, and selected those textbooks which a minimally proficient ninth-grade student could be expected to read. When the DRP unit score corresponding to those textbooks was compared to the DRP unit score of the passage identified by the first group, the scores were identical. The DRP unit score recommended by both reading sub-groups was 47 (55 items correct).

Using the same procedures, the Reading Subcommittee also recommended that a Statewide Reading Goal for Ninth-Graders be set at 60 DRP units

(69 items correct.) The ninth-grade performance goal in reading states that students should be at the instructional level for prose materials ordinarily used in the ninth grade.

At the EERA Advisory Committee meeting, a variation of those procedures was used to validate the decisions made by the Reading Subcommittee. The EERA Advisory Committee examined a passage with a DRP score close to 47. It also looked at the DRP scores of popular newspaper and magazines. The group concurred that for ninth-grade students, a score of 47 indicates the reading level of a student who could be expected to survive in the curriculum in a typical school in the state. It should be noted that this is not to say that the student with a DRP score of 47 reads well enough to comprehend all of the commonly read media publications. It is expected that the skill level of students will increase during the remaining years in high school.

As a result of these meetings, the CSDE recommended:

(1) that for those students whose DRP unit score is lower than 47 (raw score of 55), remedial assistance be given in reading; and

(2) that for those students whose DRP unit score is between 47 and 59, (raw scores between 55 and 68) the reading and academic performance of these students be reviewed and special attention be given to reading instruction within the general curriculum.

State Board Approval. In January 1980, the State Board of Education approved the standard-setting process and all four of the proposed

Statewide Levels of Expected Performance: 62 percent for Mathematics, 58 percent for Basic Writing Skills in the Language Arts, a holistic score of 4 for Writing, and a raw score of 55 items correct for Reading (47 DRP units).

The Judged Student Proficiency Status Study (JSPSS)

In order to validate the SLOEPs set in January 1980 prior to the test administration of the test, the CSDE conducted a Judged Student Proficiency Status Study (JSPSS) in March, 1980. This was designed to secure teacher judgements of approximately 4,500 students which would be used in the Contrasting and Borderline Groups analyses. This study provided the CSDE with data to examine the degree of congruence between teacher judgment and student performance on the tests. Teachers' judgments about students could be compared with teachers' judgments about items. (Hambleton and Eignor, 1978).

Contrasting and Borderline Groups Methods. Each building participating in the proficiency test had a coordinator designated by the superintendent. In March, 1980 during the week of testing, that test coordinator had the responsibility of completing the Student Proficiency Status Roster (Figure 2). The first task was to select a random sample of 30 ninth grade students who would be taking the test. The instructions for selecting the sample are outlined in Figure 2. The second task was to secure teacher judgment about the Student Sample. To accomplish this, the roster was circulated to the Mathematics and English Teachers in the school, accompanied by a list of the skills to be measured on the test (with illustrative examples). The Mathematics teachers rated the

Figure 2.

INSTRUCTIONS FOR COMPLETING THE
STUDENT PROFICIENCY STATUS ROSTER

To the Test Coordinator:

List on the reverse side the names of 30 students selected by a random sampling method for the Student Proficiency Status Validation Study. After the four test sessions have been completed at your school, circulate this roster along with a copy of the Skills and Examples booklet among the math and English teachers who teach ninth-grade students. Each student should be rated by the one teacher most familiar with that student's achievement in each subject area.

Selecting a random sample: Divide the number of ninth-grade students who are taking the EERA Proficiency Test by 30. This quotient should then be used as the interval between students on an alphabetical roster. For example, if there are 190 ninth-graders taking the test, 190 ÷ 30 = 6.3. Now start with the sixth name and select every sixth student from an alphabetical roster.

To the Teacher:

By rating the proficiency level of each student listed on the reverse side, you will be providing important information to the State Department of Education for evaluating the appropriateness of statewide levels of expected performance (cut-off scores) which have been set for each part of the Ninth-Grade Proficiency Test. The ratings will have no bearing upon the students' scores or their eligibility for remedial assistance and will be kept confidential. When you receive this roster and a copy of the Skills and Examples booklet, please proceed as follows:

(1) Read carefully the portion of the Skills and Examples booklet for your subject area. English teachers will rate students on both the Reading test and the Writing Skills in the Language Arts test. English teachers are encouraged to consult with the Reading Specialist in your school about the reading ability of certain students. Math teachers will rate students on the Mathematics test.

(2) Find the names of the students who are currently in your classes and whose work you know well.

(3) Keeping in mind the particular skills included on the test, rate each student as Proficient, Borderline, or Nonproficient. If you do not know a student well enough to make this judgment, leave this space blank. Do not use the Borderline category for these students.

- PROFICIENT: This student has general mastery of the basic skills measured on the EERA test and does not need remedial assistance beyond that provided within the general instructional program.
- BORDERLINE: This student falls within a borderline category between proficient and nonproficient and cannot be confidently assigned to either group. It is difficult to decide whether this student requires remedial assistance beyond that provided within the general instructional program.
- NONPROFICIENT: This student does not currently have general mastery of the basic skills measured on the EERA test and does need remedial assistance beyond that provided within the general instructional program.

Keep in mind that you are attempting to judge PROFICIENCY ONLY. While such issues as behavior problems, absenteeism, and attitudes may affect the grades students receive in school, they may not necessarily affect the student's proficiency rating.

In addition, it is possible that all 30 students could fall into one category. Rate each student as an individual and do not be concerned about the number of students falling in any one category.

Figure 2 (continued)

STUDENT PROFICIENCY STATUS ROSTER

Please read the instructions on the reverse side. Then rate the students listed below using the following code:

1 = Proficient (P)
2 = Borderline (B)
3 = Nonproficient (N)

Circle one number for each student for each subject area.

Student Name	Consider the basic skills being assessed by the EERA Ninth-Grade Test:								
	A. READING			B. WRITING			C. MATH		
	(P)	(B)	(N)	(P)	(B)	(N)	(P)	(B)	(N)
1.	1	2	3	1	2	3	1	2	3
2.	1	2	3	1	2	3	1	2	3
3.	1	2	3	1	2	3	1	2	3
4.	1	2	3	1	2	3	1	2	3
5.	1	2	3	1	2	3	1	2	3
6.	1	2	3	1	2	3	1	2	3
7.	1	2	3	1	2	3	1	2	3
8.	1	2	3	1	2	3	1	2	3
9.	1	2	3	1	2	3	1	2	3
10.	1	2	3	1	2	3	1	2	3
11.	1	2	3	1	2	3	1	2	3
12.	1	2	3	1	2	3	1	2	3
13.	1	2	3	1	2	3	1	2	3
14.	1	2	3	1	2	3	1	2	3
15.	1	2	3	1	2	3	1	2	3
16.	1	2	3	1	2	3	1	2	3
17.	1	2	3	1	2	3	1	2	3
18.	1	2	3	1	2	3	1	2	3
19.	1	2	3	1	2	3	1	2	3
20.	1	2	3	1	2	3	1	2	3
21.	1	2	3	1	2	3	1	2	3
22.	1	2	3	1	2	3	1	2	3
23.	1	2	3	1	2	3	1	2	3
24.	1	2	3	1	2	3	1	2	3
25.	1	2	3	1	2	3	1	2	3
26.	1	2	3	1	2	3	1	2	3
27.	1	2	3	1	2	3	1	2	3
28.	1	2	3	1	2	3	1	2	3
29.	1	2	3	1	2	3	1	2	3
30.	1	2	3	1	2	3	1	2	3

(more over)

students on their mathematics proficiency; the English teachers rated the students on both their Language Arts and Reading proficiency levels. The English teachers were encouraged to consult with the Reading Specialists in the school about the reading ability of certain students.

The teachers were asked to select those students who were currently in their classes and whose work they knew well and make one of three judgments about those students: Proficient, Borderline, or Non-Proficient. The definitions of those categories are outlined in Figure 2. The teachers were told not to use the Borderline Group for students they did not know well; rather they were told to leave the judgment blank. Teachers were also told to keep in mind that they were to judge PROFICIENCY ONLY. ("While such issues as behavior problems, absenteeism and attitudes may affect the grades students receive in school, they may not necessarily affect the student's proficiency ratirg.") Teachers were also told that it was possible that all 30 students might fall into one category.

Analyzing the Results. The judgment categories were Proficient, Borderline, and Non-Proficient. Three separate score distributions were generated, one for each group. The Contrasting Group method uses only the distributions of the Proficient and Non-Proficient groups. These data can be used in a variety of ways. The most sophisticated use of the data incorporates decision-theory and often Bayesian statistics. The various stakeholders first establish a desirable ratio of Type I and Type II errors.

This results from responses to the following two questions:

1. Which is a more serious error, a False Master or a False Non-Master?
2. How much more serious is it?

(The implications of each of these errors are discussed more fully in a later section of this paper.) A ratio is developed to express the relative error tolerance (Livingston, 1980). This ratio is then applied to the distributional data to determine the most appropriate standard.

A less sophisticated and more commonly used procedure is to plot the data for the two groups, smooth the curves and then select as a standard the point at which the two distributions overlap. (See Zieky and Livingston, 1977 p. 9 for a description of this technique.) Hill (1979) has suggested that a problem with using the overlapping distributions is that "because of random fluctuations due to small sample size, the distributions overlap at many points, and thus the decision as to which point to select can seem rather arbitrary." He suggests that one way to avoid this problem is to select as the cut off point the score that minimizes the total number of incorrect decisions. He notes that for each possible standard setting score, some students who have been judged to be competent will score below the cutoff; others who have been judged to be not competent will score at or above the cutoff. The objective

of the approach advocated by Hill would be to minimize the total number of errors.

In the tables which compare the various standard-setting methods, the Contrasting Groups technique advocated by Hill has been employed. Each of the tables presents the score which minimizes the total number of errors. In each case, the total percent of students who are misclassified at that standard is also included.

For the calculation of the Borderline Groups score, the department followed the very simple technique described by Zieky and Livingston (1977, p. 8). The median score of the Borderline group is the standard. Zieky and Livingston point out:

> If the scores of the borderline-group are spread widely over the range of possible scores (i.e., some with scores near the bottom and some with scorers near the top), then the method is not working well.

It can be seen that the state distribution of borderline students was spread over the whole range of possible scores. The reasons presented by Zieky and Livingston are:

1. The borderline group may include many students who were put in the group, not because their achievement was actually borderline but because their achievement was difficult for the teachers to judge. (These might be students who have trouble expressing themselves or who are uncooperative.)

2. The teachers may be basing their judgments on something other than what the test measures.

The instructions designed for the JSPSS were intended to encourage teachers to make their judgments on non-extraneous aspects of basic

skills proficiency. However, it is not possible to know (without a set of follow-up interviews) whether the instructions were successful.

Tables 1, 2, 3, and 4 illustrate the impact on the percent of students above and below the Mathematics, Language Arts, Writing Sample, and Reading standards respectively, when those standards are established through different standard-setting methods. Each Table revealed the following common patterns:

1. No two standard-setting methods yielded the same standard. In fact, the ranges in the percent of students scoring below the various standards were quite large.
2. The Borderline Method for setting standards always yielded the highest standard.
3. While it is true that the Borderline Method produces the highest standard, it (almost by definition) also yields the lowest percent of False Masters in each subject area.
4. When examining the Judged Student Proficiency Status Study (JSPSS) sample, it is clear, across each of the tables, that teachers have a better idea of who is proficient than who is non-proficient. The percent of students judged proficient but scoring below the standard (False non-masters) ranged from a low of only 1.8% in mathematics to a higher of 3.9% in language arts. On the other hand, the percent of students judged non-proficient but scoring at or above the standard (False Masters) ranged from a low of 49.1% in language arts to a high of 66.6% in Writing.

TABLE 1

IMPACT ON THE PERCENT OF STUDENTS ABOVE AND BELOW
THE MATHEMATICS STANDARD, WHEN SET BY DIFFERENT METHODS

STANDARD SETTING METHOD	STANDARD*	TOTAL POPULATION	JUDGED STUDENT PROFICIENCY STATUS STUDY (JSPSS)			
		% Scoring Below Standard	% Scoring Below Standard	% Judged Proficient But Scoring Below Standard (False Non-Masters)	% Judged Borderline But Scoring Below Standard	% Judged Non-Proficient But Scoring At Or Above Standard (False Masters)
		N= 44207	N= 4554	N= 3008 (66.0% of JSPSS)	N= 1137 (25.0% of JSPSS)	N=409 (9.0 % of JSPSS)
ANGOFF (N= 9)	38	16.5	12.3	3.6	21.1	47.7
NEDELSKY (N= 9)	42	23.8	19.0	7.7	32.0	34.7
CONTRASTING GROUPS (Minimum % of Total Errors = 8.5)	35	11.9	8.6	1.8	14.2	57.2
BORDERLINE GROUP (Median Score)	47	35.5	30.7	16.2	50.5	17.6
STATEWIDE LEVEL OF EXPECTED PERFORMANCE (SLOEP)**	41	21.8	17.1	6.4	29.4	37.7

*Number of Items = 65 (The STANDARD is expressed in raw score units)
Mean % Correct = 75.85 (49.30 items)
Standard Deviation = 17.27% (11.23 items)
Kuder-Richardson Reliability = .932
Standard Error of Measurement = 4.50% (2.93 items)

** SLOEP = an average of Angoff and Nedelsky

TABLE 2
IMPACT ON THE PERCENT OF STUDENTS ABOVE AND BELOW
THE LANGUAGE ARTS STANDARD, WHEN SET BY DIFFERENT METHODS

STANDARD SETTING METHOD	STANDARD*	TOTAL POPULATION	JUDGED STUDENT PROFICIENCY STATUS STUDY (JSPSS)			
		% Scoring Below Standard	% Scoring Below Standard	% Judged Proficient But Scoring Below Standard (False Non-Masters)	% Judged Borderline But Scoring Below Standard	% Judged Non-Proficient But Scoring At Or Above Standard (False Masters)
		N = 44207	N = 4554	N = 3008 (60.3% of JSPSS)	N = 1137 (27.5 of JSPSS)	N = 409 (12.2% of JSPSS)
ANGOFF (N=10)	21	8.0	5.6	.8	6.9	74.0
NEDELSKY (N=11)	18	4.8	3.4	.4	3.8	82.3
CONTRASTING GROUPS (Minimum % of Total Errors =11.5)	26	18.8	14.4	3.9	22.0	49.1
BORDERLINE GROUP (Median Score)	30	37.3	33.4	16.5	50.9	22.4
STATEWIDE LEVEL OF EXPECTED PERFORMANCE (SLOEP)**	21	8.0	5.6	.8	6.9	74.0

*Number of Items = 36 (The STANDARD is expressed in raw score units)
Mean % Correct = 82.24 (29.61 items)
Standard Deviation = 15.80% (5.69 items)
Kuder-Richardson Reliability = .875
Standard Error of Measurement - 5.60% (2.02 items)
**SLOEP = ANGOFF

TABLE 3
IMPACT ON THE PERCENT OF STUDENTS ABOVE AND BELOW
THE WRITING SAMPLE STANDARD, WHEN SET BY DIFFERENT METHODS

STANDARD SETTING METHOD	STANDARD*	TOTAL POPULATION	JUDGED STUDENT PROFICIENCY STATUS STUDY (JSPSS)			
		% Scoring Below Standard	% Scoring Below Standard	% Judged Proficient But Scoring Below Standard (False Non-Masters)	% Judged Borderline But Scoring Below Standard	% Judged Non-Proficient But Scoring At Or Above Standard (False Masters)
		N= 43533	N= 4526	N= 2741 (60.6% of JSPSS)	N= 1238 (27.4% of JSPSS)	N= 547 (12.1% of JSPSS)
SLOEP and Contrasting Groups (Minimum % of Total Errors = 14.0%)	4	11.4	9.5	3.5	12.4	66.6
BORDERLINE GROUP (Median Score)	5	32.1	29.9	16.2	41.9	28.5

*Range of Scores= 2 to 8 (The STANDARD is expressed in holistic score units).
Mean holistic score= 5.3

TABLE 4

IMPACT ON THE PERCENT OF STUDENTS ABOVE AND BELOW
THE READING STANDARD, WHEN SET BY DIFFERENT METHODS

STANDARD SETTING METHOD	STANDARD*	TOTAL POPULATION	JUDGED STUDENT PROFICIENCY STATUS STUDY (JSPSS)			
		% Scoring Below Standard	% Scoring Below Standard	% Judged Proficient But Scoring Below Standard (False Non-Masters)	% Judged Borderline But Scoring Below Standard	% Judged Non-Proficient But Scoring At Or Above Standard (False Masters)
		N= 44063	N= 4544	N= 3147 (69.3% of JSPSS)	N= 1005 (21.1% of JSPSS)	N= 392 (8.6 % of JSPSS)
STATEWIDE LEVEL OF EXPECTED PERFORMANCE (SLOEP)**	55	6.9	5.1	1.1	8.7	71.9
Statewide Goal	69	17.3	13.8	4.6	26.2	44.6
Contrasting Groups (minimum % of Total Errors = 8.1%)	63	11.8	9.2	2.4	16.1	54.1
Borderline Group	79	31.6	27.1	13.3	50.5	22.2

*Range = 0 to 98 items (The STANDARD is expressed in raw score units)
Mean Score = 81.8 items

**The SLOEP = expert judgment of the difficulty level of reading passages and textbooks that a minimally competent student would be expected to understand. See pp.9-10 for a fuller description of the procedures.

5. It appears that the sample of students in the JSPSS is above the total population in ability since the percent scoring below the standard is lower in all subject areas.

Table 5 illustrates the percent of students passing and failing the two language arts subtests by the total population and the JSPSS sample. Two patterns are evident from the table. First, the JSPSS sample is slightly above the total population in ability. Second, English teachers in the JSPSS have a better idea of who is proficient than who is non-proficient. But perhaps the power of the table is revealed by the column labelled "Judged Non-Proficient." These are the False Masters. When both parts of the language arts test (writing sample and multiple choice test) are taken into account rather than viewed in isolation, the classification error for false masters (56.3%) is greatly reduced. This is because some students fail one part of the language arts test but pass the other. The classification error in Table 2 for the language arts false masters is 74.0% while the writing sample false masters classification error in Table 3 is 66.6%.

TABLE 5

PERCENT OF STUDENTS PASSING AND FAILING THE TWO LANGUAGE ARTS SUBTESTS

PASS/FAIL COMBINATION ON TWO LANGUAGE ARTS SUBTESTS	TOTAL POPULATION	JUDGED STUDENT PROFICIENCY STATUS STUDY (JSPSS)			
		Judged Sample (JSPSS)	Judged Proficient	Judged Borderline	Judged Non-Proficient
	N= 43681	N= 4551	N= 2720 (60.4% of JSPSS)	N= 1249 (27.4% of JSPSS)	N= 552 (12.1% of JSPSS)
Pass Language Arts / Pass Writing Sample	84.8	87.3	95.7	82.4	56.3
Pass Language Arts / Fail Writing Sample	7.7	7.3	3.5	10.7	18.3
Fail Language Arts / Pass Writing Sample	3.2	2.6	.4	4.2	9.8
Fail Language Arts / Fail Writing Sample	4.4	2.8	.4	2.6	15.6

DISCUSSION

Policy Issues

There are several issues policy makers must consider when setting standards. These relate to the purposes and penalties of the test and the ability of the standard-setters to adjust the standard over time. These will be discussed in turn.

Purposes and Penalties. If a test is designed to identify that group of students who may have fallen through the cracks of earlier identification mechanisms and who are in need of remedial assistance, there is the temptation to set a standard which reduces the number of False Masters, that is, the number of students in need of help who "pass the test." In this way, virtually everyone who needs help will be identified. However, there are "penalties" associated with such a policy. First, it will increase the number of students who are False Non-Masters, that is, students who really do have mastery level skills but who fall below the standard. This may have a negative impact on (1) a student's academic self-concept, (2) on teachers' expectations for that student (e.g., Pygmalion effect) and (3) on peer reactions. In addition, the school district must be ready to face the wrath of parents who have been told that their children who have been attaining school grades B or C do not demonstrate basic skills proficiency. Also the financial costs in personnel time and diagnostic testing/assessment will increase relative to the level of the standard. Lastly, there is the public perception of the educational system. In an attempt to identify virtually every student who may need assistance, many

students will be falsely identified. By overrepresenting the number of non-masters, the perception of the public will be that schools are not doing an adequate job in preparing youngsters to master the basic skills. In an educational climate affected by over a decade of declining SAT scores and thirty-seven states already on the basic skills testing bandwagon, this is an important consideration. Do we want an artificially high standard to further darken the black eye that has been given to public education?

Looking at the impact of the complementary situation, setting a relatively low standard, one sees quickly the major advantage of such a policy--it decreases the number of false non-masters. That is, only the very poorest students will fall below the standard. This avoids many of the problems outlined above. Relatively speaking, fewer students will be misidentified as non-masters, fewer families will be upset, the costs to schools will be reduced, and the public confidence in education will be less shaken. But one disadvantage looms large--this practice increases the number of False Masters, i.e., students who really need remediation but pass the test. If the real purpose of the legislation or board policy was to identify students who have been slipping through the cracks of the system, this situation provides a veritable Catch - 22.

In addition, when proficiency standards are part of a high school graduation requirement the stakes for students and school districts become quite high. On such a examination, can the system afford to identify large

numbers of False Non-Masters? But conversely, can the system afford to set the standard too low? After all, the policy of having the test in place probably resulted from pressure exerted by a variety of dissatified constituencies (e.g., business and industry, institutions of higher learning, parents) who felt that the high school diploma has lost its meaning. To set standards low allows a large number of False-Masters and defeats the purpose of the policy.

Flexible Standards. Can standards be set with the understanding that they may be changed given a change in circumstances? It would appear that this is a reasonable and sensible approach. But will the public allow a high standard to be lowered? If a standard is resulting in too many false Non-masters will the public understand and support the lowering of the standard? Or, will the reaction be one of suspicion? Will newspaper headlines criticize educators for reducing the problem by hiding it? In contrast, the public might be willing to allow standards to be raised, particularly if students' scores improve over time. The possibility of public tolerance for flexibility in only one direction provides some help to standard-setters: One should obviously always attempt to approximate the *true or real or best* standard which reduces the incidence of False-Masters and False Non-Masters. However, in a world where tests are imperfectly valid and reliable and where human performance at any point in time is influenced by physical, attitudinal, motivational and environment factors as well as ability and skill level, no one cut score will ever do all

things for all people. It may be pragmatic to choose the lower standard and increase it over time.

Policy Implications

Table 6 illustrates that not only did the different standard setting techniques yield different cut scores, but they also resulted in different identified problem areas. The Contrasting Groups technique and the percent of students in the Judged Student Proficiency Status Study judged non-proficient identify Basic Writing Skills in the Language Arts as the most serious problem area while a combination of Angoff and Nedelsky highlights Mathematics as the problem area. The Borderline Group Method illustrates that all three areas are of critical importance.

TABLE 6

PERCENT OF TOTAL POPULATION SCORING BELOW THE STANDARD USING DIFFERENT STANDARD-SETTING METHODS WITH ACCOMPANYING POLICY IMPLICATIONS

METHOD	PERCENT OF STUDENTS SCORING BELOW STANDARD				IDENTIFIED PROBLEM AREA
	READING (R)	LANGUAGE ARTS (LA)	WRITING SAMPLE (WS)	MATHEMATICS (M)	
% of JSPSS JUDGED NON-PROFICIENT	8.6	12.2	12.1	9.0	LA & WS
CONTRASTING GROUPS	11.8	18.8	11.4	11.9	LA
BORDERLINE GROUP	31.6	37.3	32.1	35.5	ALL
SLOEP (Combinations of Angoff and Nedelsky)	6.9	8.0	11.4	21.8	M

LIMITATIONS OF STANDARD-SETTING METHODS

Setting Statewide Versus Local Standards.

Many of the standard-setting procedures were originally established for use in relatively homogeneous locations. Defining the minimally competent student in a local school district is relatively concrete. Teachers in a school district know the students, the curriculum, the expectations of parents and local employers from the surrounding communities. They can discuss the differences between setting standards based on the present versus the desired state of affairs. In applying a method in which it is necessary to imagine or to simulate the behavior of a group of minimally proficient students, they could even agree on specific students in the school who would serve as the prototype. Similarly, when making judgments about the performance of actual students, they would be using the same students on whom to make judgments. The degree of congruence among the different methods should be greater than when one tries to apply these methods at the state level.

When setting standards, the first task is that of defining minimal proficiency. Teachers from different school systems have different conceptions of minimal proficiency. For the standard to apply to all districts, the standard-setters must envision the typical district in the state. Then, within the typical district, they must imagine a group of minimally proficient students. There are two levels of abstraction necessary. Trying to conjure up an image of the typical district is hard enough, but adding to that the notion of the minimally proficient student within that typical district makes the task even harder. That is what

standard-setters must hold in mind continuously when using either Angoff or Nedelsky methods at the state level.

Differences Between Angoff and Nedelsky Methods

One of the major differences between the Angoff and Nedelsky methods is in the task to be done. The Angoff group makes a global determination of the percentage of minimally proficient students who would answer the question correctly. The Nedelsky group approximates the behavior of students who solve the question by using a process of elimination. Clearly, not all students use such a process. In fact, it may be that only the students who perform relatively well use such a technique, and only for the test items for which they are not certain.

The person using the Angoff method may feel that an item is quite easy and assign to it a probability of .85 or .90. The same person using the Nedelsky method could never arrive at such an estimate. The only possible probabilities for an item using the Nedelsky method are .25, .33, .50 and 1.00. The absence of the middle range probabilities may account for some of the difference resulting from the various methods.

The Limitations of the Present JSPSS

There are three important limitations of the Contrasting Groups and Borderline Group methods as employed in the present study.* The first of these is related to the sampling plan employed in the present study. In order to ease the reporting burden of local district personnel, our JSPSS sampled thirty students in each of the approximately 200 buildings in which testing took place. This sampling plan had the effect of under-

*The largest drawback of the Contrasting and Borderline Groups methods was described earlier in this paper--the fact that teachers often judge students who are not performing adequately as non-proficient. As Zieky and Livingston pointed out, there may be a large variety of explanations for poor performance, only one of which is low proficiency.

TABLE 7

SIZE OF POPULATION AND JUDGED STUDENT PROFICIENCY STATUS SAMPLE AND PROFICIENCY TEST SCORES BY TYPE OF COMMUNITY

TYPE OF COMMUNITY	TOTAL POPULATION		JUDGED STUDENT PROFICIENCY STATUS SAMPLE (JSPSS)		PROFICIENCY TEST SCORES FOR TOTAL POPULATION							
					MATHEMATICS		LANGUAGE ARTS		WRITING		READING	
	N	%	N	%	Total: Mean % Correct	% At or Above SLOEP	Total: Mean % Correct	% At or Above SLOEP	Mean Holistic Score	% At or Above SLOEP	Mean Total Score	% At or Above SLOEP
Large City (1)	6251	14.31	404	8.88	60.2	43.6	68.7	72.8	4.4	71.8	68.6	77.0
Fringe City (2)	9867	22.59	976	21.45	79.6	85.1	85.5	96.0	5.5	92.5	84.8	96.2
Medium City (3)	10376	23.75	823	18.08	76.3	80.2	83.0	93.2	5.3	90.4	82.6	94.7
Suburban Town (4)	7921	18.13	934	20.52	81.2	88.5	86.9	97.2	5.7	94.0	86.2	97.2
Emerging Suburban (5)	3373	7.72	493	10.83	79.4	85.6	85.5	96.2	5.6	92.3	84.5	96.7
Rural Town (6)	3118	7.14	503	11.05	76.9	81.8	83.6	94.3	5.4	90.5	83.2	94.7
Vocational-Technical Schools	2775	6.35	418	9.18	76.3	83.6	80.5	95.0	4.8	84.3	80.8	95.8
State	43681	100.0	4551	100.0	75.8	78.2	82.2	92.0	5.3	88.6	81.8	93.1

representing students in large buildings. Because large buildings tend to be found in large school districts and because the large school districts tended to perform less well on the state tests, this sampling plan resulted in an underrepresentation of students who performed less well on the test. Table 7 displays the percentage of students in the total population and in the JSPSS in each type of community (TOC). The performance of each TOC is also provided to demonstrate the underrepresentation of students who scored low in the JSPSS. In future studies of this type, a different sampling plan would be used. For example, the JSPSS might consider using 10% of the students in each school. (The only difficulty with such a plan is that some schools only test 50 or 60 students and to sample only 5 or 6 of those students may not be enough.)

The second limitation of the JSPSS is that the definition of proficiency used by teachers was different from that used by the Angoff and Nedelsky groups. This was done intentionally because it was felt that teachers would have a less difficult time making judgments about students if they used their own (or locally determined) definition of proficiency. If the sample had been more representative of the total state, these differences would have averaged out. But, because the poorer students were underrepresented in the study, it is more important to acknowledge these definitional differences. If more teachers used higher standards in making judgments, the standard would be affected. Specifically, the judged Non-proficient and Borderline groups would be composed of more able students which would identify more False Masters. Furthermore, the median of the Borderline Group and the point of overlapping curves of the Contrasting Group would both be higher.

A final drawback of the Contrasting and Borderline Groups methods as applied here is that English teachers may, in reality, have insufficient information on which to assess their student's reading ability.

RECOMMENDATIONS

The issues and problems addressed in this paper are both technical and political in nature. Since it is generally recognized that all standard setting methods rely on human judgement, therefore rendering them subjective, it may not be possible to develop completely satisfactory solutions to these problems. The recommendations which follow are provided to sensitize policy makers and standard setters to the primary issues and provide some possible solutions.

1. Standards derive their meaning from the methods used to set them. For this very reason, the standard setting method and its procedures must be carefully selected and well documented.
2. Table 6 illustrated that different standard setting techniques results in different cut scores and problem areas. For this reason it is best to use more than one standard setting method so that the implications of each cut score can be observed and analyzed.
3. It is important to understand that there are two types of possible standard setting errors, one of which may be far more costly than the other. Cost may be in financial dollars psychological harm, public confidence, and teacher behavior. A district or state attempting to establish a standard for a graduation requirement may have a different cost-benefit ratio from one merely trying to identify students who may be in need of remedial assistance. This understanding will provide important guidance to the standared setters.

4. Further research is needed to examine the subject area interaction with standard setting techniques.
5. Setting state level standards is both a technical and political issue. A few of the technical issues were discussed in the LIMITATIONS section. To avoid political pitfalls, involve all of the stakeholders in the standard-setting process (Shepard, 1978). That is, use a broad spectrum of interest groups and constituencies to (1) determine which method(s) to use and (2) to set the standard. One reason for doing so is that they may provide valuable imputs. Secondly, these groups must absorb the impact of the standards and be able to accept the consequences of their judgments. The standard-setters will be much less vulnerable later if all of the major constituencies have been involved.
6. When utilizing the Contrasting and Borderline Group methods to set state level standards carefully design the sampling plan employed to select students. It is very important not to underrepresent a particular segment of the test taking student population.

7. Consider having a common core of people set the standard for all parts of the test. Most often, mathematics experts set the mathematics standard and language arts experts set the language arts score. Even when using the same method differences in the standards may in part arise from difference in the groups of people setting the standards in addition to differences between the tests.

8. Spend an appropriate amount of time defining minimally proficient. This definition should be related to the purpose of the test--e.g., identification for remedial assistance or graduation from high school. All standard-setters should attend the definitional meeting. Questions will arise related to who will be taking the test, (special education, bilingual, etc.) Furthermore, the standard-setters should be clear as to whether they are setting standards to mirror the *real* or the *ideal* state of affairs. Specifically, a discussion should occur which specifies whether the standard is to reflect what current minimal performance *is* or what desired or ideal minimal performance *should be*. The former is less risky than the later but may offend some teachers.

REFERENCES

Andrew, B.J. & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.H. Scales, norms and equivalent scores. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C. American Council on Education. 1971, 514-515.

Angoff, W.H. A Technique for the Investigation of Cultural Differences. Paper presented at the meeting of the American Psychological Association, Honolulu, May, 1972.

Behuniak, P. An Investigation of Standard Setting Methods for Objective-Referenced Proficiency Tests. Dissertation, The University of Connecticut. 1980.

Glass, G.V. Minimum competence and incompetence in Florida. Kappan, 1978, 59, 602-205.

Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R.K. & Eignor, D. Determining Minimum Competency Levels: A Consideration of Issues, Methods and Implementation Strategies. Paper presented at the American Educational Research Association Conference on Minimal Competency Testing, Washington, D.C. 1978.

Hill, R. Suggestions for the Interpretation of Results Using the Contrasting Groups Method. A discussion paper prepared for The Massachusetts State Department of Education, 1979. RMC Corporation.

Koffler, S.A. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 77, 167-178.

Livingston, S.A. Using Decision Theory to Establish Cutoff Scores on Tests. Seminar held at Educational Testing Service, Princeton, New Jersey, February, 1980.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Shepard, L. Setting standards. In Gunda and Sanders (Eds.) Practices and Problems in Competency Based Measurement. National Council of Measurement in Education, 1979, 59-71.

Zieky, M.L. & Livingston, S.A. Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, New Jersey, Educational Testing Service, 1977.